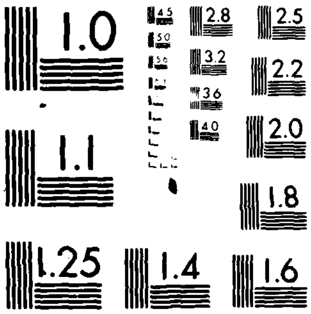
AD-A112 469 ILLINOIS UNIV AT CHICAGO CIRCLE DEPT OF QUANTITATIVE --ETC F/G 12/1
MULTI-SAMPLE CLUSTER ANALYSIS USING AKAIKE'S INFORMATION CRITER--ETC(U)
JAN 82 H BOZDOGAN, S L SCLOVE N00014-80-C-0408
UNCLASSIFIED TR-82-1 NL

1 OF 1
AD
A112469

END
DATE
FILMED
4-82
DTIC


1.0
4.5
5.0
5.6
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

College of Business Administration, University of Illinois at Chicago Circle

UICC

1

# MULTI-SAMPLE CLUSTER ANALYSIS USING AKAIKE'S INFORMATION CRITERION*

by

HAMPARSUM BOZDOGAN
Quantitative Methods Department
University of Illinois at Chicago Circle

and

STANLEY L. SCLOVE
Departments of Mathematics and Quantitative Methods
University of Illinois at Chicago Circle

TECHNICAL REPORT NO. 82-1
January 30, 1982

PREPARED FOR THE
OFFICE OF NAVAL RESEARCH
UNDER
CONTRACT N00014-80-C-0408,
TASK NR042-443
with the University of Illinois at Chicago Circle

Principal Investigator: Stanley L. Sclove

QUANTITATIVE METHODS DEPARTMENT
UNIVERSITY OF ILLINOIS AT CHICAGO CIRCLE
CHICAGO, ILLINOIS 60680

*Presented by the first author as an Invited Paper, Special Session on Cluster Analysis, 789th Meeting, American Mathematical Society, University of Massachusetts, Amherst, MA, October 16-18, 1981.

VV45

# MULTI-SAMPLE CLUSTER ANALYSIS USING AKAIKE'S INFORMATION CRITERION*

Hamparsum Bozdogan and Stanley L. Sclove
University of Illinois at Chicago Circle

## CONTENTS

*Presented by the first author as an Invited Paper, Special Session on Cluster Analysis, 789th Meeting, American Mathematical Society, University of Massachusetts, Amherst, MA, October 16-18, 1981.

# MULTI-SAMPLE CLUSTER ANALYSIS USING AKAIKE'S INFORMATION CRITERION*

Hamparsum Bozdogan and Stanley L. Sclove
University of Illinois at Chicago Circle

## ABSTRACT

Multi-sample cluster analysis, the problem of grouping samples, is studied from an information-theoretic viewpoint via Akaike's Information Criterion (AIC). This criterion combines the maximum value of the likelihood with the number of parameters used in achieving that value. The multi-sample cluster problem is defined, and AIC is developed for this problem.

The form of AIC is derived in both univariate and multivariate analysis of variance models. Numerical examples are presented and results are shown to demonstrate the utility of AIC in identifying the best clustering alternatives.

Key Words and Phrases: Multi-sample cluster analysis; Akaike's Information Criterion (AIC); ANOVA Model, MANOVA Model; maximum likelihood.

*Presented by the first author as an Invited Paper, Special Session on Cluster Analysis, 789th Meeting, American Mathematical Society, University of Massachusetts, Amherst, MA, October 16-18, 1981.

# MULTI-SAMPLE CLUSTER ANALYSIS
# USING AKAIKE'S INFORMATION CRITERION*

Hamparsum Bozdogan and Stanley L. Sclove
University of Illinois at Chicago Circle

## 1. Introduction

In this paper, we shall develop Akaike's Information Criterion (AIC) for multi-sample cluster analysis. The problem of multi-sample cluster analysis arises when we are given a collection of samples (groups, treatments), to be clustered into homogeneous groups.

It is reasonable to provide a practically useful statistical procedure that would use some sort of statistical model to aid in comparisons of various collections of samples, identify homogeneous groups of samples, and tell us which should be clustered together and which samples should not.

Examples of multi-sample clustering situations are abundant. Here we mention a few.

Example 1.1. Botany: grouping of three types of species of iris, namely *Iris setosa* (S), *Iris versicolor* (Ve), and *Iris virginica* (Vi), given in Example 6.2 and Table 6.3 in Section 6, on the basis of each and of all the four variables.

Example 1.2. Zoology: grouping of geographical locations to study the differences of populations of two types of species of *Crocidura*. Delany and Healy [7] studied variation in white-toothed shrews, that is, nocturnal mammals, in the British Isles. White-toothed shrews of genus *Crocidura* occur in the Channel and Scilly Islands of the British Isles and the French mainland. From $p = 10$ measurements on each of $n = 399$ skulls obtained from the $K = 10$ locations, Tresco, Bryher, St. Agnes, St. Martin's, St. Mary's, Sark, Jersey, Alderney, Guernsey, and Cap Gris Nez. The sample sizes for the data from the

*Presented by the first author as an Invited Paper, Special Session on Cluster Analysis, 789th Meeting, American Mathematical Society, University of Massachusetts, Amherst, MA, October 16-18, 1981.

ten locations are, respectively, $n_1 = 144$, $n_2 = 16$, $n_3 = 12$, $n_4 = 7$, $n_5 = 90$, $n_6 = 25$, $n_7 = 6$, $n_8 = 26$, $n_9 = 53$, $n_{10} = 20$. Attempts were made to analyze the pattern of variation between these ten populations to examine the belief that there may be two species of Crocidura, namely Crocidura russula, Crocidura suaveolens. The locations were geographically close, but it is assumed that only one sub-species was present in any one place. Thus the problem here is to cluster the locations, that is, "samples" into homogeneous groups to discover the origin of the two species.

Example 1.3. Air and Water Pollution: grouping of weather class types or nitrate sites to distinguish whether the source of nitrate is weather type or local. Heidorn [12] studied synoptic, that is, general weather patterns associated with nitrates in southern Ontario. In recent years, there has been growing concern over the potential hazard of particulate nitrate in the atmosphere which acts as a respiratory irritant, especially to those who have asthma problems. Nitrate is also suspected to lower the pH level in freshwater lakes.

A sample of $n = 17$ cities across southern Ontario from Windsor in the west to Kingston in the east was chosen as the location of nitrate sites. Nitrate concentrations for the 17 sites were measured. In order to determine the effect of weather patterns on the measurement of nitrate, eight weather class types were defined for the nitrate sites. Thus the problem here is to cluster the weather class types or the sites into homogeneous groups to determine whether the source of particulate nitrate is due to weather class type or is local.

Example 1.4. Business and Economics: grouping of corporations by their financial characteristics. Chen et al. [6], Williams and Goodman [16], and others, studied the statistical methods for clustering corporations on the basis of yearly data concerning several of their financial characteristics. Thus the general problem here is to cluster the sets of corporations in order to detect, describe and distinguish relatively homogeneous groups of companies so that the formation of the groups and organizational behavior of companies can be studied and compared.

So, as we see, multi-sample cluster analysis examples are quite rich and varied.

The analysis of variance (ANOVA) is a widely used model for comparing two or more univariate samples, where the familiar Student's t and F statistics are used for formal comparisons among two or more samples. Multivariate analysis of variance (MANOVA) is a widely used model for comparing two or more multivariate samples. In the MANOVA model, the likelihood ratio principle leads to Wilks' [17] lambda, or in short Wilks' $\Lambda$ criterion as the test statistic. It plays the same role in multivariate analysis that the F-ratio statistic plays in the univariate case.

Often, however, the formal analyses involved in MANOVA are not revealing or informative. Therefore, in this paper we shall propose Akaike's Information Criterion (AIC) as a new procedure for comparing the clusters, and use it to identify the best clustering alternatives.

In 1971, Akaike first introduced an information criterion, referred to as an automatic (model) identification criterion or Akaike's information criterion (AIC), for the identification and comparison of statistical models in a class of competing models with different numbers of parameters. It is defined by

$$\text{(1.1)}\quad \text{AIC} = -2 \log_e (\text{maximized likelihood}) \\ +2\,(\text{number of independently adjusted parameters within the model}).$$

It was obtained with the aid of an information theoretic interpretation of the method of maximum likelihood by Akaike ([2], [3]). It estimates minus twice the expected log likelihood of the model whose parameters are determined by the method of maximum likelihood. When several competing models are being compared or fitted, AIC is a simple procedure which measures the badness of fit or the discrepancy of the estimated model from the true model when a set of data is given.

The first term in (1.1) stands for the penalty of badness of fit or downward bias when the maximum likelihood estimators of the parameters of the model are used. The second term in the definition of AIC, on the other hand, stands for the penalty of increased unreliability or compensation for the bias in the first term as a consequence of increasing number of parameters. If more parameters are used to describe the data, it is natural to get a larger likelihood, possibly without improving the true goodness of fit by penalizing the use of additional parameters.

Thus, when there are several competing models, the parameters within the models are estimated by the method of maximum likelihood and the AIC-values are computed and compared to find a model with the minimum value of AIC. This procedure is called the minimum AIC procedure. The model with the minimum AIC is called the minimum AIC estimate (MAICE) and is designated as the best model.

In Section 2, we shall define the general multi-sample cluster problem, and in Section 3, we shall briefly discuss the number of clustering alternatives for a given K groups or samples into k nonempty clusters. In the subsequent sections, that is, in Section 4 and in 5, we shall derive the AIC procedure for the univariate analysis of variance (ANOVA) model, and the multivariate analysis of variance (MANOVA) model. In Section 6, we shall give numerical examples for both univariate and multivariate multi-sample cluster analysis on real data sets to demonstrate our results of AIC and minimum AIC procedures obtained from different computer analyses.

## 2. The Multi-Sample Cluster Problem

Suppose each individual, object, or case, has been measured on p response or outcome measures (dependent variables) simultaneously in K independent groups or samples (factor levels). Let

$$\underline{X}\,(n \times p) = \begin{bmatrix} \underline{X_1} \\ \underline{X_2} \\ \cdot \\ \cdot \\ \cdot \\ \underline{X_K} \end{bmatrix} \tag{2.1}$$

be a single data matrix of K groups or samples, where $\underline{X}_g$ $(n_g x p)$ represents the observations from the g-th group or sample, g=1,2,...,K, and n $\sum_{g=1}^{K} n_g$. The goal of cluster analysis is to put the K groups or samples into k homogeneous groups, samples, or classes where k is unknown, but $k \leq K$.

Often individuals or objects have been sampled from K>1 populations. For multi-samples or multiple groups of individuals or objects the data matrix may be represented in partitioned form as above. Let $n_g$ represent the number of individuals in the g-th (random) sample, g=1,2,...,K. The $n_g$ are not restricted to being equal or proportional to other $n_g$'s. The total number of observations is $n = \sum_{g=1}^{K} n_g$. Let $X_{gi}$ be the px1 vector of observations in group g=1,2,...,K, and for individual i=1,2,...,$n_g$.

## 3. The Number of Clustering Alternatives for a Given K Samples into k Nonempty Clusters

In this section, we shall briefly discuss how to obtain the total number of clustering alternatives for a given K, the number of groups or samples. For this, we shall recall some established results.

Theorem 3.1. The number of ways of clustering K groups or samples into k clusters such that none of the k clusters is empty is given by

$$(3.1) \qquad \sum_{g=0}^{k} \binom{k}{g} (-1)^g \, (k-g)^K \; ,$$

where the order of groups or samples within each cluster is irrelevant.

Proof. Duran and Odell [9].

In this theorem the k clusters are assumed to be distinct. However, in clustering K groups or samples into k clusters, none of which is empty, the order of the k clusters is irrelevant. Consequently, from this fact and Theorem 3.1, it follows that the total number of ways of clustering K groups or samples into k clusters is given by

$$(3.2) \qquad S(K,k) = \frac{1}{k!} \sum_{g=0}^{k} \binom{k}{g} \, (-1)^g \, (k-g)^K$$

which is known as the Stirling Number of the Second Kind (see, e.g., Abramowitz and Stegun [1]) and also called the number of clustering alternatives.

If k, the number of clusters of groups or samples is known in advance, then the total number of clustering alternatives is given by S(K,k). However, if k is not specified a priori and varies, then the total number of clustering alternatives for a given K, the number of groups or samples, is given by

$$(3.3) \qquad \sum_{k=1}^{K} S(K,k) \quad .$$

Table 3.1 gives S(K,k) for values of K and k up to 10.

TABLE 3.1. NUMBER OF CLUSTERING ALTERNATIVES FOR VARIOUS VALUES OF K AND k

| k | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| K | | | | | | | | | | | |
| 1 | 1 | | | | | | | | | | 1 |
| 2 | 1 | 1 | | | | | | | | | 2 |
| 3 | 1 | 3 | 1 | | | | | | | | 5 |
| 4 | 1 | 7 | 6 | 1 | | | | | | | 15 |
| 5 | 1 | 15 | 25 | 10 | 1 | | | | | | 52 |
| 6 | 1 | 31 | 90 | 65 | 15 | 1 | | | | | 203 |
| 7 | 1 | 63 | 301 | 350 | 140 | 21 | 1 | | | | 877 |
| 8 | 1 | 123 | 966 | 1701 | 1050 | 266 | 28 | 1 | | | 4136 |
| 9 | 1 | 255 | 3021 | 7770 | 6951 | 2645 | 462 | 36 | 1 | | 21142 |
| 10 | 1 | 511 | 9318 | 34101 | 42525 | 22821 | 5879 | 750 | 45 | 1 | 115952 |

Consider, for example, K=3 samples. We now wish to cluster K=3 groups or samples first into k=3 groups or samples, then into k=2 groups or samples, and k=1 group or sample in a hierarchical fashion.

From Table 3.1, we have the total number of ways of clustering K=3 groups or samples into k=3 homogeneous groups or samples is 1. The total number of ways of clustering K=3 groups or samples into k=2 homogeneous groups or samples is 3. The total number of ways of clustering k=3 groups or samples into k=1 homogeneous group or sample is 1. Thus adding up these results, we obtain, in total 5 clustering alternatives as the total for K=3 groups or samples into k=1,2, and 3 homogeneous groups. We note that 5 is nothing but the sum of the values of row 3 in Table 3.1.

The 5 clustering alternatives can be classified according to their representation forms to make it easy to list all 5 possible clustering alternatives. The representation forms in this case are denoted by

(i) {1} {1} {1},

(ii) {2} {1},

(iii) {3},

where each of the components in a representation {g} denotes the number, g, of groups or samples in the corresponding cluster. The components of a representation form will always be written in a hierarchical order to depict the patterns of clustering alternatives. In our example there are 5 clustering alternatives but only 3 representation forms. In general the number of representation forms is much smaller then the number of clustering alternatives.

We now list the clustering alternatives corresponding to their representation forms in Table 3.2 as follows:

TABLE 3.2. A SIMPLE PATTERN OF CLUSTERING ALTERNATIVES WHEN K=3 AND k=3, 2, and 1

| Alternatives | Clustering | Number of Parameters m |
|---|---|---|
| 1 | (1) (2) (3) | 3 |
| 2 | (1 2) (3) | 2 |
| 3 | (1 3) (1) | 2 |
| 4 | (2 3) (1) | 2 |
| 5 | (1 2 3) | 1 |

For example, in alternative one, the group or sample 1, 2, and 3 are clustered as singletons. In terms of a hypothesis on means, this corresponds to $\mu_1$, $\mu_2$, and $\mu_3$ all being different, and therefore, the number of parameters, m, is equal to 3. Hence, indicating that group 1, 2, and 3 are all heterogeneous. In alternative two, groups or samples 1 and 2 are clustered together forming a homogeneous subset, and group or sample 3 is clustered alone forming a heterogeneous subset. In terms of a hypothesis on means, this corresponds to $\mu_1 = \mu_2$, and $\mu_3$ is different from both $\mu_1$ and $\mu_2$ with the total number of parameters m being equal to 2. In a similar fashion, we interpret the other clustering alternatives continuing down the line of Table 3.2.

As a last example, we shall just list the results of the total number of possible clustering alternatives when K=4 groups or samples in Table 3.3 as follows.

TABLE 3.3. A SIMPLE PATTERN OF CLUSTERING ALTERNATIVES WHEN K=4 AND k=4, 3, 2, AND 1

| Alternatives | Clustering | Number of Parameters, m |
|---|---|---|
| 1 | (1) (2) (3) (4) | 4 |
| 2 | (1 2) (3) (4) | 3 |
| 3 | (1 3) (2) (4) | 3 |
| 4 | (1 4) (2) (3) | 3 |
| 5 | (2 3) (1) (4) | 3 |
| 6 | (2 4) (1) (3) | 3 |
| 7 | (3 4) (1) (2) | 3 |
| 8 | (1 2) (3 4) | 2 |
| 9 | (1 3) (2 4) | 2 |
| 10 | (1 4) (2 3) | 2 |
| 11 | (1 2 3) (4) | 2 |
| 12 | (1 2 4) (3) | 2 |
| 13 | (1 3 4) (2) | 2 |
| 14 | (2 3 4) (1) | 2 |
| 15 | (1 2 3 4) | 1 |

In concluding this section, we see that in general the total number of ways of clustering K groups or samples into k homogeneous groups or samples is given by equation (3.2), and the total number of possible clustering alternatives is given by the expression (3.3).

## 4. AIC For The Univariate Model

We now turn our attention to consider situations with several univariate normal samples. The general layout for such data (one-way ANOVA) is represented in the following tabular form.

TABLE 4.1. GENERAL DATA REPRESENTATION FOR ONE-WAY ANOVA

| | Groups | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | . . . | K | |
| Observations | $z_{11}$ | $z_{21}$ | . . . | $z_{K_2}$ | |
| | $z_{12}$ | $z_{22}$ | . . . | $z_{K_2}$ | |
| | . | . | . . . | . | |
| | . | . | . . . | . | |
| | . | . | . . . | . | |
| | $z_{1n_1}$ | $z_{2n_2}$ | . . . | $z_{Kn_K}$ | |
| TOTALS | $T_1$ | $T_2$ | . . . | $T_K$ | $T$ |
| SAMPLE SIZES | $n_1$ | $n_2$ | . . . | $n_K$ | $n = \sum_{g=1}^{K} n_g$ |
| SAMPLE MEANS | $\bar{z}_{1\cdot}$ | $\bar{z}_{2\cdot}$ | . . . | $\bar{z}_{K\cdot}$ | $\bar{\bar{z}}$ |
| VARIANCES | $s_1^2$ | $s_2^2$ | | $s_K^2$ | $s^2$ |

For example, we may have multi-sample data with samples of sizes $n_1, n_2, \ldots, n_K$ which are assumed to have come from K populations, the first with mean $\mu_1$ and variance $\sigma^2$, the second with mean $\mu_2$ and variance $\sigma^2, \ldots,$ the Kth with mean $\mu_K$ and variance $\sigma^2$. We may want to compare these K group or sample means $\mu_1, \mu_2, \ldots, \mu_K$ given that all have a common $\sigma^2$. Hence, this is the well known analysis of variance (ANOVA) model. In terms of the parameters the ANOVA model is $\underset{\sim}{\theta} = (\mu_1, \mu_2, \ldots, \mu_K, \sigma^2)$ with m=k+1 parameters, where k is the number of groups.

We shall derive the form of AIC for this model. Recall the definition of AIC from Section 1,

$$\begin{aligned} \text{AIC} &= -2 \log_e L(\hat{\underset{\sim}{\theta}}) + 2m \\ &= -2 \log_e (\text{maximized likelihood}) + 2m , \end{aligned}$$

where m denotes the number of independently adjusted parameters within the model.

Suppose there are K independent samples of independent observations, with $n_g$, g=1,2,...,K, observations in the g-th group and $n = \sum_{g=1}^{K} n_g$. Denote the unknown means of the groups by $\mu_1, \mu_2, \ldots, \mu_K$. Assume that the samples $(z_{11}, z_{12}, \ldots, z_{1n_1}; \ldots; z_{K1}, \ldots, z_{Kn_K})$ are drawn randomly from K populations which are $N(\mu_g, \sigma^2)$. If the groups can differ only in their means, we may express this as

(4.1) $z_{gi} = \mu_g + \varepsilon_{gi}$, $g=1,2,\ldots,K$; $i=1,2,\ldots,n_g$,

where $z_{gi}$ is the value of the response or outcome variable in the g-th group for the i-th individual or object,

$\mu_g$ are parameters,

$\varepsilon_{gi}$ are independent $N(0,\sigma^2)$ error variables.

This equation is called the one-way ANOVA model.

Thus, the basic null hypothesis of interest in this case is given by

$$(4.2) \qquad H_0 : \mu_1 = \mu_2 = \cdot\cdot\cdot\cdot = \mu_K.$$

The alternative hypothesis is given by

$H_1$ : the K population means are not all equal.

Every analysis of variance involves a partitioning of the total sum of squares of deviations, SST, into the within-group sum of squares of deviations, SSW, and the between-group sum of squares of deviations, SSB. For more details on this, we refer the reader to any basic text on statistics, e.g., Anderson and Sclove [4].

We now derive the form of Akaike's Information Criterion (AIC) for the one-way ANOVA model given in (4.1).

The likelihood function is given by

$$(4.3) \qquad L(\{\mu g\},\sigma^2;\underset{\sim}{z}) = (2\pi\sigma^2)^{-n/2}\exp[-\sum_{g=1}^{K}\sum_{i=1}^{ng}(z_{gi} - \mu g)^2/(2\sigma^2)].$$

The log likelihood function is

$$(4.4) \qquad l(\{\mu g\},\sigma^2;\underset{\sim}{z}) \equiv \log L(\{\mu g\},\sigma^2;\underset{\sim}{z})$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{g=1}^{K}\sum_{i=1}^{ng}(z_{gi} - \mu_g)/(2\sigma^2).$$

As is well known, the MLE's are

$$(4.5) \qquad \hat{\mu}_g = \frac{1}{n_g}\sum_{i=1}^{n_g} z_{gi} = \bar{z}_{g.},\ g=1,2,\ldots,K,$$

and

$$(4.6) \qquad \hat{\sigma}^2 = \frac{1}{n} \sum_{g=1}^{K} \sum_{i=1}^{n_g} (z_{gi} - \bar{z}_{g.})^2 = \frac{SSW}{n} ,$$

where $SSW = \sum_{g=1}^{K} \sum_{i=1}^{n_g} (z_{gi} - \bar{z}_{g.})^2$, the Within Group Sum of Squares.

Substituting these back into (4.4), we have

$$\begin{aligned} l(\{\hat{\mu}_g\},\hat{\sigma}^2;\underset{\sim}{z}) &\equiv \log L(\{\hat{\mu}\},\hat{\sigma}^2;\underset{\sim}{z}) \\ &= -\frac{n}{2}[\log(2\pi) + \log \frac{SSW}{n}] - \frac{n}{2} . \end{aligned}$$

Since

$$(4.7) \qquad AIC = -2 \log_e L(\hat{\underset{\sim}{\theta}}) + 2m,$$

where m is the number of parameters, and since

$$(4.8) \qquad -2 \log L(\{\hat{\mu}_g\},\hat{\sigma}^2) = n \log(2\pi) + n \log \frac{SSW}{n} + n ,$$

then AIC becomes

$$(4.9) \qquad AIC = n \log(2\pi) + n \log \frac{SSW}{n} + n + 2(k+1).$$

Since the constants do not affect the result of comparison of models, we could ignore them and use the simplified version

$$(4.10) \qquad AIC^* = n\log_e SSW + 2(k+1)$$

where $\qquad n = \sum_{g=1}^{K} n_g$ = the total sample size, SSW = Within Group Sum of Squares, and

k = number of groups or samples compared, or the number of independently adjusted parameters within the model.

However, for purposes of comparison we retain the constants and use AIC.

## 5. AIC For the Multivariate Model

In this section we shall study the natural extension of the univariate model we considered in Section 4 to its multivariate analogue. Therefore, throughout this section we shall suppose that we may have independent data matrices $\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_K$, where the rows of $\underline{X}_g$ $(n_g x p)$ are independent and identically distributed (i.i.d.) $N_p(\underset{\sim}{\mu}_g, \underline{\Sigma})$, g=1,2,...,K. In terms of the parameters $\underset{\sim}{\theta} = (\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \ldots, \underset{\sim}{\mu}_K, \underline{\Sigma})$ the model we shall consider here is

$$\underset{\sim}{\theta} = (\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \ldots, \underset{\sim}{\mu}_K, \underline{\Sigma})$$

with m = kp + p(p+1)/2 parameters, where k is the number of groups, and p is the number of variables.

As in the univariate case, consider K normal populations with different mean vectors $\underset{\sim}{\mu}_g$, g=1,2,...,k,...,K. Let $\underset{\sim}{z}_{gi}$, g=1,2,...,K; i=1,2,...,$n_g$, be a random sample of observations from the g-th population $N_p(\underset{\sim}{\mu}_g, \underline{\Sigma})$. If the groups or samples can differ only in their mean vectors, we can write the multivariate one-way analysis variance (MANOVA) model as

$$(5.1) \qquad \underset{\sim}{z}_{gi} = \underset{\sim}{\mu}_g + \underset{\sim}{\varepsilon}_{gi} \ , \ g=1,\ldots,K; \ i=1,2,\ldots,n_g \ ,$$

where $\underset{\sim}{z}_{gi}$ is the (p x 1) response or outcome vector in the g-th group for i-th individual or object,

$\underset{\sim}{\mu}_g$ are vector parameters, and

$\underset{\sim}{\varepsilon}_{gi}$ are independent $N_p(\underset{\sim}{0}, \underline{\Sigma})$ random vector errors.

Thus, the basic null hypothesis we usually are interested in testing is given by

$$(5.2)\qquad H_0 : \underset{\sim}{\mu}_1 = \underset{\sim}{\mu}_2 = \ldots = \underset{\sim}{\mu}_K .$$

The alternative hypothesis is given by

$$H_1 \ : \text{Not all } \underset{\sim}{\mu}_K \text{ are equal.}$$

Wilks' lambda is a general statistic for handling this problem. Although, there are several other conventional statistics for this purpose, they all can be viewed as special cases of Wilks' $\Lambda$ which we shall not discuss here.

For notational purposes, we shall denote $\underline{T}$ to be the "total" sum of squares and products (SSP) matrix, $\underline{W}$ to be the "within-group" or "within-sample" SSP matrix, and $\underline{B}$ to be the "between-group" SSP matrix. Hence, it can be shown that

$$(5.3)\qquad \underline{T} = \underline{W} + \underline{B} ,$$

where

$$(5.4)\qquad \underline{T} = \sum_{g=1}^{K} \sum_{i=1}^{n_g} (\underset{\sim}{z}_{gi} - \underset{\sim}{\bar{z}})(\underset{\sim}{z}_{gi} - \underset{\sim}{\bar{z}})',$$

$$(5.5)\qquad \underline{W} = \sum_{g=1}^{K} \sum_{i=1}^{n_g} (\underset{\sim}{z}_{gi} - \underset{\sim}{\bar{z}}_g)(\underset{\sim}{z}_{gi} - \underset{\sim}{\bar{z}}_g)',$$

and

$$(5.6)\qquad \underline{B} = \sum_{g=1}^{K} n_g \, (\underset{\sim}{\bar{z}}_g - \underset{\sim}{\bar{z}})(\underset{\sim}{\bar{z}}_g - \underset{\sim}{\bar{z}})',$$

with

$$\bar{\underset{\sim}{z}}_g = \frac{1}{n_g}\sum_{i=1}^{n_g} \underset{\sim}{z}_{gi} \ , \ g=1,2,\ldots,K \ ,$$

$$\bar{\underset{\sim}{z}} = \frac{1}{n}\sum_{g=1}^{K}\sum_{i=1}^{n_g} \underset{\sim}{z}_{gi} \ , \ n = \sum_{g=1}^{K} n_g \ .$$

Therefore, we can present multivariate one-way analysis of variance (MANOVA) table as follows.

TABLE 5.1. MANOVA TABLE

| Source | d.f. | SSP matrix | Wilks' criterion |
|---|---|---|---|
| Between samples | K-1 | $\underline{B}$ | $\dfrac{\lvert\underline{W}\rvert}{\lvert\underline{T}\rvert}$ |
| Within samples | n-K | $\underline{W}$ | $\sim\Lambda(p \ ; \ n-K \ ; \ K-1)$ |
| Total | n-1 | $\underline{T}$ | |

Now, we derive the form of Akaike's Information Criterion (AIC) for the MANOVA model given in (5.1), subject to the constraint given in (5.2). The likelihood function of all the sample observations is given by

$$(5.7) \qquad L(\mu_g,\underline{\Sigma}_g;\underline{Z}) = \prod_{g=1}^{K} L_g(\mu_g,\underline{\Sigma}_g;\underline{Z}_g),$$

or by

$$(5.8) \qquad L = (2\pi)^{-np/2} \prod_{g=1}^{K} |\underline{\Sigma}_g|^{-n_g/2} x$$

$$\exp \{-1/2 \mathrm{tr} \sum_{g=1}^{K} \underline{\Sigma}_g^{-1} \underline{A}_g - 1/2 \mathrm{tr} \sum_{g=1}^{K} n_g \underline{\Sigma}_g^{-1} (\bar{\underline{z}}_g - \underline{\mu}_g)(\bar{\underline{z}}_g - \underline{\mu}_g)'\} ,$$

where $n = \sum_{g=1}^{K} n_g$ and $\underline{A}_g = \sum_{i=1}^{n_g} (\underline{z}_{gi} - \bar{\underline{z}}_g)(\underline{z}_{gi} - \bar{\underline{z}}_g)'$ .

The log likelihood function is

$$(5.9) \qquad l(\underline{\mu}_g, \underline{\Sigma}; \underline{Z}) \equiv \log_e L$$

$$= -\frac{np}{2} \log(2\pi) - 1/2 \sum_{g=1}^{K} n_g \log|\underline{\Sigma}_g| - 1/2 \mathrm{tr} \sum_{g=1}^{K} \underline{\Sigma}_g^{-1} \underline{A}_g$$

$$- 1/2 \mathrm{tr} \sum_{g=1}^{K} n_g \underline{\Sigma}_g^{-1} (\bar{\underline{z}}_g - \underline{\mu})(\bar{\underline{z}}_g - \underline{\mu}_g)' .$$

Since the common covariance matrix is $\underline{\Sigma}$, then the log likelihood function becomes

$$(5.10) \qquad l\{\underline{\mu}_g\}, \underline{\Sigma}; \underline{Z}) \equiv \log_e L(\{\underline{\mu}_g\}, \underline{\Sigma}; \underline{Z})$$

$$= -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log |\underline{\Sigma}| - 1/2 \mathrm{tr} \underline{\Sigma}^{-1} \sum_{g=1}^{K} \underline{A}_g$$

$$- 1/2 \mathrm{tr} \underline{\Sigma}^{-1} \sum_{g=1}^{K} n_g (\bar{\underline{z}}_g - \underline{\mu}_g)(\bar{\underline{z}}_g - \underline{\mu}_g)' ,$$

and the maximum-likelihood estimates (MLE's) of $\underset{\sim}{\mu}_g$, and $\underline{\Sigma}$ are

(5.11) $$\hat{\underset{\sim}{\mu}}_g = \bar{\underset{\sim}{z}}_g \text{ , } g=1,2,\ldots,K,$$

and

(5.12) $$\hat{\underline{\Sigma}} = n^{-1}\underline{W},$$

where $$\underline{W} = \sum_{g=1}^{K} \underline{A}_g \text{ .}$$

Substituting these back into (5.10) and simplifying, the maximized log likelihood becomes

(5.13) $$l(\{\hat{\underset{\sim}{\mu}}_g\},\hat{\underline{\Sigma}};\underline{Z}) \equiv \log L(\{\hat{\underset{\sim}{\mu}}_g\},\hat{\underline{\Sigma}};\underline{Z})$$

$$= -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|n^{-1}\underline{W}| - \frac{np}{2} \text{ ,}$$

where $\underline{W}$ is the "within-group" SSP matrix.

Since

(5.14) $$\text{AIC} = -2 \log_e L(\hat{\underset{\sim}{\theta}}) + 2m \text{ ,}$$

where $m = kp + \frac{p(p+1)}{2}$ is the number of parameters, then AIC becomes

(5.15) $$\text{AIC} = np\log(2\pi) + n\log|n^{-1}\underline{W}| + np + 2[kp + \frac{p(p+1)}{2}].$$

Since the constants do not affect the result of comparison of models, we could ignore them and reduce the form of AIC to a much simpler form

(5.15) $$\text{AIC}^* = n\log_e|\underline{W}| + 2[kp + \frac{p(p+1)}{2}]$$

where $n = \sum_{g=1}^{K} n_g$ = the total sample size,

$|\underline{W}|$ = the determinant of "within-group" SSP matrix,

$k$ = number of groups or samples compared,

$p$ = number of variables.

However, for purposes of comparison we retain the constants and use AIC.

## 6. Numerical Examples of Multi-Sample Cluster Analysis on Real Data Sets

In this section we shall give numerical examples of both univariate and multivariate multi-sample data, and cluster the groups or samples, and choose the best clusterings by using Akaike's Information Criterion (AIC) as derived in Sections 4 and 5.

Our computations were carried out for all the examples we shall present here on an IBM 370, using various statistical software packages such as MINITAB, SPSS, and SPEAKEASY (VM/CMS version).

### 6.1. Univariate Examples

For the univariate numerical examples we shall illustrate our results on two data sets, a biomedical data set of Dolkart, Halpern, and Perlman [8] and Fisher [10] iris data. Here we shall take 150 iris specimens on each of the four morphological variables: sepal length and width and petal length and width and demonstrate our results on these variables individually rather than considering all of them together.

Example 6.1. (Brown and Hollander [5]) Antibody Responses in Three Groups of Mice: "Dolkart, Halpern, and Perlman [8] compared antibody responses in normal and alloxan diabetic mice. Their investigation was designed to study the circulating antibody response in alloxan diabetic, insulin-treated diabetic and normal CF-1 mice injected with serum albumin.

"Only those animals treated with alloxan who had elevated serum glucose levels (250mg/100 ml or higher) were included in the study, together with a group of normal animals. Animals were bled from the orbital sinus, and the serum analyzed for antigen binding capacity of BSA, glucose concentration, and serum proteins. BSA was iodinated with I-131, and the antigen-binding capacity of each serum sample was determined as micrograms of BSA nitrogen bound by 1 ml of undiluted serum." The data are given in Table 6.1.

TABLE 6.1 MICROGRAMS OF BSA NITROGEN BOUND PER ml OF UNDILUTED MOUSE SERUM ON DAY 39, FOLLOWING INJECTION OF 5 mg BSA ANTIGEN INTO EACH ANIMAL ON DAY 0 AND 28

| Normal | Alloxan Diabetic | Alloxan Diabetic-Treated with Insulin |
|---|---|---|
| 155.76 | 390.72 | 82.50 |
| 282.00 | 46.20 | 99.66 |
| 197.34 | 468.60 | 97.66 |
| 297.00 | 86.46 | 150.48 |
| 115.50 | 174.02 | 242.88 |
| 126.72 | 132.66 | 67.98 |
| 119.46 | 13.20 | 227.70 |
| 29.04 | 498.96 | 130.68 |
| 252.78 | 167.64 | 73.26 |
| 122.10 | 62.04 | 17.82 |
| 349.14 | 127.38 | 19.80 |
| 108.90 | 275.88 | 100.32 |
| 143.22 | 176.22 | 71.94 |
| 64.02 | 145.86 | 133.32 |
| 25.54 | 108.24 | 464.64 |
| 85.80 | 275.88 | 36.96 |
| 122.10 | 50.16 | 46.20 |
| 454.85 | 72.60 | 34.32 |
| 655.38 | | 43.56 |
| 13.86 | | |

Source: R.E. Dolkart, B. Halpern, and J. Perlman [8].

In this example we are given K=3 groups or samples and we wish to cluster them into k=1, 2, and 3 homogeneous groups. From Table 3.1, as we know, there are 5 total possible clustering alternatives, namely, (1) (2) (3) all separate, and (1 2) (3), (1 3) (2), (2 3) (1), and (1 2 3) all together. Let us code Normal Group=1, Alloxan Diabetic Group=2, and Alloxan Diabetic-Treated with Insulin Group=3. Considering the ANOVA model as our underlying model for comparisons of these groups, from a simple ANOVA run on the computer we computed the AIC's for each of the 5 clustering alternatives. The results are shown in Table 6.2.

TABLE 6.2 THE AIC'S FOR ANTIBODY RESPONSES IN THREE GROUPS OF MICE

| Alternative | Clustering | $n\log_e(2\pi)$ | $n\log_e {}^{SSW}/_n$ | n | k | 2(k+1) | AIC |
|---|---|---|---|---|---|---|---|
| 1 | (1) (2) (3) | 104.758 | 559.139 | 57 | 3 | 8 | 728.897[c] |
| 2 | (1 2) (3) | 104.758 | 559.149 | 57 | 2 | 6 | 726.907[a] |
| 3 | (1 3) (2) | 104.758 | 561.945 | 57 | 2 | 6 | 729.703 |
| 4 | (2 3) (1) | 104.758 | 561.513 | 57 | 2 | 6 | 729.271 |
| 5 | (1 2 3) | 104.758 | 562.581 | 57 | 1 | 4 | 728.339[b] |

$n = 20 + 18 + 19 = 57$

$AIC = n\log_e(2\pi) + n\log_e {}^{SSW}/_n + n + 2\,(k+1)$

[a] First Minimum AIC

[b] Second Minimum AIC

[c] Third Minimum AIC

In this example the first minimum AIC occurs at the alternative submodel 2. That is, the MAICE is submodel 2 indicating to us that in terms of clustering, Normal Group=1 and Alloxan Diabetic Group=2 should be clustered together, and Alloxan Diabetic-Treated with Insulin Group=3 should be clustered by itself. Therefore, in terms of a hypothesis on means, (1 2) (3) corresponds to $\mu_1 = \mu_2 \neq \mu_3$ indicating that Normal and Alloxan Diabetic Groups form the best homogeneous set in terms of their nitrogen-binding capacities, and the Alloxan Diabetic-Treated with Insulin Group forms a set by itself. On the other hand, the second minimum AIC occurs at the alternative submodel 5, and the third minimum AIC is at the alternative submodel 1 indicating that either we should cluster all the groups together or treat each group separately, but if we were to compare each group separately to the Normal Group=1, then we should choose Normal Group=1 with Alloxan Diabetic Group=2 together as the best choice by the minimum AIC procedure.

*Example 6.2*. *Clustering of Irises by Groups*: As we mentioned in Example 1.2, the iris data set is composed of 150 iris species belonging to three groups or species, namely *Iris setosa* (S), *Iris versicolor* (Ve), and *Iris virginica* (Vi) measured on sepal and petal length and width. Each group is represented by 50 plants. The data set for the 150 irises are given in Table 6.3.

This data set has been quite extensively studied in classification and cluster analysis since it was published by Fisher [10], and still today, is being used as a "testing ground" for classification and clustering methods proposed by many investigators such as Friedman and Rubin [11], Kendall [13], Solomon [15], Mezzich and Solomon [14], and many others, including the present authors.

For each of the 150 plants we already know the group structure of the iris species, namely K=3 groups or samples. Even though the two species, *Iris setosa* and *Iris versicolor* were found growing in the same colony, and *Iris virginica* was found growing in a different colony, Fisher reports in his linear discriminant analysis the separation of *I. setosa* completely from *I. versicolor* and *I. virginica*. Since then other investigators have shown similar results in their studies such as the ones we mentioned above.

TABLE 6.3 . MEASUREMENTS ON THREE TYPES OF IRIS

| *Iris setosa* | | | | *Iris versicolor* | | | | *Iris virginica* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sepal length | Sepal width | Petal length | Petal width | Sepal length | Sepal width | Petal length | Petal width | Sepal length | Sepal width | Petal length | Petal width |
| 5.1 | 3.5 | 1.4 | 0.2 | 7.0 | 3.2 | 4.7 | 1.4 | 6.3 | 3.3 | 6.0 | 2.5 |
| 4.9 | 3.0 | 1.4 | 0.2 | 6.4 | 3.2 | 4.5 | 1.5 | 5.8 | 2.7 | 5.1 | 1.9 |
| 4.7 | 3.2 | 1.3 | 0.2 | 6.9 | 3.1 | 4.9 | 1.5 | 7.1 | 3.0 | 5.9 | 2.1 |
| 4.6 | 3.1 | 1.5 | 0.2 | 5.5 | 2.3 | 4.0 | 1.3 | 6.3 | 2.9 | 5.6 | 1.8 |
| 5.0 | 3.6 | 1.4 | 0.2 | 6.5 | 2.8 | 4.6 | 1.5 | 6.5 | 3.0 | 5.8 | 2.2 |
| 5.4 | 3.9 | 1.7 | 0.4 | 5.7 | 2.8 | 4.5 | 1.3 | 7.6 | 3.0 | 6.6 | 2.1 |
| 4.6 | 3.4 | 1.4 | 0.3 | 6.3 | 3.3 | 4.7 | 1.6 | 4.9 | 2.5 | 4.5 | 1.7 |
| 5.0 | 3.4 | 1.5 | 0.2 | 4.9 | 2.4 | 3.3 | 1.0 | 7.3 | 2.9 | 6.3 | 1.8 |
| 4.4 | 2.9 | 1.4 | 0.2 | 6.6 | 2.9 | 4.6 | 1.3 | 6.7 | 2.5 | 5.8 | 1.8 |
| 4.9 | 3.1 | 1.5 | 0.1 | 5.2 | 2.7 | 3.9 | 1.4 | 7.2 | 3.6 | 6.1 | 2.5 |
| 5.4 | 3.7 | 1.5 | 0.2 | 5.0 | 2.0 | 3.5 | 1.0 | 6.5 | 3.2 | 5.1 | 2.0 |
| 4.8 | 3.4 | 1.6 | 0.2 | 5.9 | 3.0 | 4.2 | 1.5 | 6.4 | 2.7 | 5.3 | 1.9 |
| 4.8 | 3.0 | 1.4 | 0.1 | 6.0 | 2.2 | 4.0 | 1.0 | 6.8 | 3.0 | 5.5 | 2.1 |
| 4.3 | 3.0 | 1.1 | 0.1 | 6.1 | 2.9 | 4.7 | 1.4 | 5.7 | 2.5 | 5.0 | 2.0 |
| 5.8 | 4.0 | 1.2 | 0.2 | 5.6 | 2.9 | 3.6 | 1.3 | 5.8 | 2.8 | 5.1 | 2.4 |
| 5.7 | 4.4 | 1.5 | 0.4 | 6.7 | 3.1 | 4.4 | 1.4 | 6.4 | 3.2 | 5.3 | 2.3 |
| 5.4 | 3.9 | 1.3 | 0.4 | 5.6 | 3.0 | 4.5 | 1.5 | 6.5 | 3.0 | 5.5 | 1.8 |
| 5.1 | 3.5 | 1.4 | 0.3 | 5.8 | 2.7 | 4.1 | 1.0 | 7.7 | 3.8 | 6.7 | 2.2 |
| 5.7 | 3.8 | 1.7 | 0.3 | 6.2 | 2.2 | 4.5 | 1.5 | 7.7 | 2.6 | 6.9 | 2.3 |
| 5.1 | 3.8 | 1.5 | 0.3 | 5.6 | 2.5 | 3.9 | 1.1 | 6.0 | 2.2 | 5.0 | 1.5 |
| 5.4 | 3.4 | 1.7 | 0.2 | 5.9 | 3.2 | 4.8 | 1.8 | 6.9 | 3.2 | 5.7 | 2.3 |
| 5.1 | 3.7 | 1.5 | 0.4 | 6.1 | 2.8 | 4.0 | 1.3 | 5.6 | 2.8 | 4.9 | 2.0 |
| 4.6 | 3.6 | 1.0 | 0.2 | 6.3 | 2.5 | 4.9 | 1.5 | 7.7 | 2.8 | 6.7 | 2.0 |
| 5.1 | 3.3 | 1.7 | 0.5 | 6.1 | 2.8 | 4.7 | 1.2 | 6.3 | 2.7 | 4.9 | 1.8 |
| 4.8 | 3.4 | 1.9 | 0.2 | 6.4 | 2.9 | 4.3 | 1.3 | 6.7 | 3.3 | 5.7 | 2.1 |
| 5.0 | 3.0 | 1.6 | 0.2 | 6.6 | 3.0 | 4.4 | 1.4 | 7.2 | 3.2 | 6.0 | 1.8 |
| 5.0 | 3.4 | 1.6 | 0.4 | 6.8 | 2.8 | 4.8 | 1.4 | 6.2 | 2.8 | 4.8 | 1.8 |
| 5.2 | 3.5 | 1.5 | 0.2 | 6.7 | 3.0 | 5.0 | 1.7 | 6.1 | 3.0 | 4.9 | 1.8 |
| 5.2 | 3.4 | 1.4 | 0.2 | 6.0 | 2.9 | 4.5 | 1.5 | 6.4 | 2.8 | 5.6 | 2.1 |
| 4.7 | 3.2 | 1.6 | 0.2 | 5.7 | 2.6 | 3.5 | 1.0 | 7.2 | 3.0 | 5.8 | 1.6 |
| 4.8 | 3.1 | 1.6 | 0.2 | 5.5 | 2.4 | 3.8 | 1.1 | 7.4 | 2.8 | 6.1 | 1.9 |
| 5.4 | 3.4 | 1.5 | 0.4 | 5.5 | 2.4 | 3.7 | 1.0 | 7.9 | 3.8 | 6.4 | 2.0 |
| 5.2 | 4.1 | 1.5 | 0.1 | 5.8 | 2.7 | 3.9 | 1.2 | 6.4 | 2.8 | 5.6 | 2.2 |
| 5.5 | 4.2 | 1.4 | 0.2 | 6.0 | 2.7 | 5.1 | 1.6 | 6.3 | 2.8 | 5.1 | 1.5 |
| 4.9 | 3.1 | 1.5 | 0.2 | 5.4 | 3.0 | 4.5 | 1.5 | 6.1 | 2.6 | 5.6 | 1.4 |
| 5.0 | 3.2 | 1.2 | 0.2 | 6.0 | 3.4 | 4.5 | 1.6 | 7.7 | 3.0 | 6.1 | 2.3 |
| 5.5 | 3.5 | 1.3 | 0.2 | 6.7 | 3.1 | 4.7 | 1.5 | 6.3 | 3.4 | 5.6 | 2.4 |
| 4.9 | 3.6 | 1.4 | 0.1 | 6.3 | 2.3 | 4.4 | 1.3 | 6.4 | 3.1 | 5.5 | 1.8 |
| 4.4 | 3.0 | 1.3 | 0.2 | 5.6 | 3.0 | 4.1 | 1.3 | 6.0 | 3.0 | 4.8 | 1.8 |
| 5.1 | 3.4 | 1.5 | 0.2 | 5.5 | 2.5 | 4.0 | 1.3 | 6.9 | 3.1 | 5.4 | 2.1 |
| 5.0 | 3.5 | 1.3 | 0.3 | 5.5 | 2.6 | 4.4 | 1.2 | 6.7 | 3.1 | 5.6 | 2.4 |
| 4.5 | 2.3 | 1.3 | 0.3 | 6.1 | 3.0 | 4.6 | 1.4 | 6.9 | 3.1 | 5.1 | 2.3 |
| 4.4 | 3.2 | 1.3 | 0.2 | 5.8 | 2.6 | 4.0 | 1.2 | 5.8 | 2.7 | 5.1 | 1.9 |
| 5.0 | 3.5 | 1.6 | 0.6 | 5.0 | 2.3 | 3.3 | 1.0 | 6.8 | 3.2 | 5.9 | 2.3 |
| 5.1 | 3.8 | 1.9 | 0.4 | 5.6 | 2.7 | 4.2 | 1.3 | 6.7 | 3.3 | 5.7 | 2.5 |
| 4.8 | 3.0 | 1.4 | 0.3 | 5.7 | 3.0 | 4.2 | 1.2 | 6.7 | 3.0 | 5.2 | 2.3 |
| 5.1 | 3.8 | 1.6 | 0.2 | 5.7 | 2.9 | 4.2 | 1.3 | 6.3 | 2.5 | 5.0 | 1.9 |
| 4.6 | 3.2 | 1.4 | 0.2 | 6.2 | 2.9 | 4.3 | 1.3 | 6.5 | 3.0 | 5.2 | 2.0 |
| 5.3 | 3.7 | 1.5 | 0.2 | 5.1 | 2.5 | 3.0 | 1.1 | 6.2 | 3.4 | 5.4 | 2.3 |
| 5.0 | 3.3 | 1.4 | 0.2 | 5.7 | 2.8 | 4.1 | 1.3 | 5.9 | 3.0 | 5.1 | 1.8 |

With this in mind, let us take K=3 groups or species on each of the variables separately and cluster them into k=1, 2, and 3 homogeneous groups. Since we are dealing with K=3 groups, by now we know that there are 5 total possible clustering alternatives. Denoting I. setosa by S, I. versicolor by Ve, and I. virginica by Vi, we have (S) (Ve) (Vi), (S, Ve) (Vi), (S, Vi) (Ve), (Ve, Vi) (S), and (S, Ve, Vi) as all the possible clustering alternatives of three iris species. Using the ANOVA model as our underlying model for comparisons of these iris groups, from a simple ANOVA run on the computer by using SPSS MANOVA program which performs both univariate and multivariate linear estimation and tests of hypotheses, we obtained the AIC's for each of the 5 clustering alternatives of iris groups on each of the four variables separately. We report our results on each of the four variables, respectively, as follows.

TABLE 6.4. THE AIC'S FOR IRISES BY GROUPS ON VARIABLE SEPAL LENGTH

| Alternative | Clustering | $n\log_e(2\pi)$ | $n\log_e SSW/n$ | n | k | 2(k+1) | AIC |
|---|---|---|---|---|---|---|---|
| 1 | (S) (Ve) (Vi) | 275.681 | -200.295 | 150 | 3 | 8 | 233.386[a] |
| 2 | (S, Ve) (Vi) | 275.681 | -135.669 | 150 | 2 | 6 | 296.012 |
| 3 | (S, Vi) (Ve) | 275.681 | - 58.550 | 150 | 2 | 6 | 373.131 |
| 4 | (Ve, Vi) (S) | 275.681 | -163.740 | 150 | 2 | 6 | 267.941[b] |
| 5 | (S, Ve, Vi) | 275.681 | - 56.966 | 150 | 1 | 4 | 372.715 |

TABLE 6.5. THE AIC'S FOR IRISES BY GROUPS ON VARIABLE SEPAL WIDTH

| Alternative | Clustering | $n\log_e(2\pi)$ | $n\log_e SSW/n$ | n | k | 2(k+1) | AIC |
|---|---|---|---|---|---|---|---|
| 1 | (S) (Ve) (Vi) | 275.681 | -326.949 | 150 | 3 | 8 | 106.732[a] |
| 2 | (S, Ve) (Vi) | 275.681 | -252.915 | 150 | 2 | 6 | 178.766 |
| 3 | (S, Vi) (Ve) | 275.681 | -287.157 | 150 | 2 | 6 | 144.524 |
| 4 | (Ve, Vi) (S) | 275.681 | -318.019 | 150 | 2 | 6 | 113.662[b] |
| 5 | (S, Ve, Vi) | 275.681 | -250.129 | 150 | 1 | 4 | 179.552 |

TABLE 6.6. THE AIC'S FOR IRISES BY GROUPS ON VARIABLE PETAL LENGTH

| Alternative | Clustering | $n\log_e(2\pi)$ | $n\log_e{}^{SSW}/_n$ | n | k | 2(k+1) | AIC |
|---|---|---|---|---|---|---|---|
| 1 | (S) (Ve) (Vi) | 275.681 | -255.988 | 150 | 3 | 8 | 177.693$^a$ |
| 2 | (S, Ve) (Vi) | 275.681 | 59.442 | 150 | 2 | 6 | 491.123 |
| 3 | (S, Vi) (Ve) | 275.681 | 163.259 | 150 | 2 | 6 | 594.940 |
| 4 | (Ve, Vi) (S) | 275.681 | -116.579 | 150 | 2 | 6 | 315.102$^b$ |
| 5 | (S, Ve, Vi) | 275.681 | 169.493 | 150 | 1 | 4 | 599.174 |

TABLE 6.7. THE AIC'S FOR IRISES BY GROUPS ON VARIABLE PETAL WIDTH

| Alternative | Clustering | $n\log_e(2\pi)$ | $n\log_e{}^{SSW}/_n$ | n | k | 2(k+1) | AIC |
|---|---|---|---|---|---|---|---|
| 1 | (S) (Ve) (Vi) | 275.681 | -478.966 | 150 | 3 | 8 | -45.285$^a$ |
| 2 | (S, Ve) (Vi) | 275.681 | -216.942 | 150 | 2 | 6 | 214.739 |
| 3 | (S, Vi) (Ve) | 275.681 | - 84.552 | 150 | 2 | 6 | 347.129 |
| 4 | (Ve, Vi) (S) | 275.681 | -314.688 | 150 | 2 | 6 | 116.993$^b$ |
| 5 | (S, Ve, Vi) | 275.681 | - 82.452 | 150 | 1 | 4 | 347.229 |

$AIC = n\log_e(2\pi) + n\log_e{}^{SSW}/_n + n + 2(k+1)$

[a]First Minimum AIC

[b]Second Minimum AIC

Looking at each of the tables above, we see that on each of the variables the first minimum AIC occurs at the alternative submodel 1, namely (S) (Ve) (Vi). That is, the MAICE is submodel 1 indicating that indeed there are three types of species across all the variables. But the second minimum AIC is at the alternative submodel 4 again across all the variables indicating that if we were to cluster any iris species, we should cluster _I. versicolor_ and _I. virginica_ together, as one homogeneous group.

Thus our minimum AIC results for each of the variables confirm other investigators' findings, including Fisher's results on the iris data. Moreover, if we were to choose among the submodels then we would choose the one with smallest minimum AIC as the best submodel. Examining the Tables 6.4, 6.5, 6.6, and 6.7, we see that the smallest minimum AIC occurs at the submodel 1 in Table 6.7 on variable petal width. This indicates that petal width alone separates the three iris species with virtual certainty, confirming again Fisher's results (see, e.g., Fisher [10]).

## 6.2. A Multivariate Example

Now we consider Fisher iris data again and this time we cluster K=3 groups or species into k=1, 2, and 3 homogeneous groups on the basis of all the four variables, assuming the MANOVA model as the underlying model for comparisons of these three iris groups. On the iris data, running SPSS MANOVA program, we obtain the following "within-group" sum of squares and products (SSP) matrices for each of the clustering alternatives. These are:

$$(1)\ (S)\ (VE)\ (VI)\qquad \underset{\sim}{W}_1 = \begin{bmatrix} 39.462 & 13.818 & 24.729 & 5.6554 \\ 13.818 & 16.962 & 8.1208 & 4.8084 \\ 24.729 & 8.1208 & 27.223 & 6.2718 \\ 5.6554 & 4.8084 & 6.2718 & 6.1566 \end{bmatrix}$$

$$150 \log_e |150^{-1} \underset{\sim}{W}_1| = -1{,}504.2$$

$$(2)\ (S,\ VE)\ (VI)\qquad \underset{\sim}{W}_2 = \begin{bmatrix} 60.714 & -1.3489 & 89.222 & 30.549 \\ -1.3489 & 27.786 & -37.906 & -12.958 \\ 89.222 & -37.906 & 222.94 & 81.818 \\ 30.549 & -12.958 & 81.818 & 35.317 \end{bmatrix}$$

$$150 \log_e |150^{-1} \underset{\sim}{W}_2| = -1{,}085.9$$

(3) (S, VI) (VE) $$\underline{W}_3 = \begin{bmatrix} 101.52 & -4.3257 & 186.38 & 76.044 \\ -4.3257 & 22.115 & -38.301 & 15.395 \\ 186.38 & -38.301 & 445.43 & 188.28 \\ 76.044 & -15.395 & 188.28 & 85.367 \end{bmatrix}$$

$$150 \log_e |150^{-1} \underline{W}_3| = -988.39$$

(4) (VE, VI) (S) $$\underline{W}_4 = \begin{bmatrix} 50.352 & 17.184 & 46.047 & 17.205 \\ 17.184 & 18.002 & 14.71 & 8.3784 \\ 46.047 & 14.71 & 68.954 & 28.882 \\ 17.205 & 8.3784 & 28.882 & 18.407 \end{bmatrix}$$

$$150 \log_e |150^{-1} \underline{W}_4| = -1{,}129.6$$

(5) (S, VE, VI) $$\underline{W}_5 = \begin{bmatrix} 102.6 & -6.0197 & 189.78 & 76.884 \\ -6.0197 & 28.307 & -49.119 & -18.124 \\ 189.78 & -49.119 & 464.33 & 193.05 \\ 76.884 & -18.124 & 193.05 & 86.57 \end{bmatrix}$$

$$150 \log_e |150^{-1} \underline{W}_5| = -941.73$$

After carrying out all our computations for each of the clustering alternatives (using the Matrix Algebra Routines in SPEAKEASY interactive computer package), we obtain the AIC's from (5.15). The results are shown in Table 6.8.

TABLE 6.8. THE AIC'S FOR IRISES BY GROUPS ON ALL VARIABLES

| Alternative | Clustering | $np\log_e(2\pi)$ | $n\log_e\lvert n^{-1}\underline{W}\rvert$ | np | k | 2m | AIC |
|---|---|---|---|---|---|---|---|
| 1 | (S) (Ve) (Vi) | 1,102.724 | -1,504.2 | 600 | 3 | 44 | 242.524[a] |
| 2 | (S, Ve) (Vi) | 1,102.724 | -1,085.9 | 600 | 2 | 36 | 652.824 |
| 3 | (S, Vi) (Ve) | 1,102.724 | - 988.39 | 600 | 2 | 36 | 750.334 |
| 4 | (Ve, Vi) (S) | 1,102.724 | -1,299.6 | 600 | 2 | 36 | 439.124[b] |
| 5 | (S, Ve, Vi) | 1,102.724 | - 941.73 | 600 | 1 | 28 | 788.994 |

$n = 150$ plants, $p = 4$ variables

$m = kp + p(p+1)/2$ parameters

$AIC = np\log_e(2\pi) + n\log_e\lvert n^{-1}\underline{W}\rvert + np + 2m$

[a] First Minimum AIC

[b] Second Minimum AIC

Hence, looking at the Table 6.8, we see that, using all four variables simultaneously the first minimum AIC occurs at the alternative submodel 1, that is, when (S) (Ve) (Vi) are all clustered separately. This indicates again that indeed there are three types of species. Therefore, the MAICE is submodel 1. Not surprisingly, the second minimum AIC occurs at the alternative submodel 4 telling us that if we were to cluster any one of the two iris groups, we should cluster _I. veriscolor_ and _I. virginica_ together as one homogeneous group, and we should cluster _I. setosa_ completely separate as one heterogeneous group.

Here, it is important to note that we obtained also the same results when we used the four variables separately in our computation of AIC in the previous section, which is encouraging.

Thus, in concluding, we see from these numerical results that AIC and consequently minimum AIC procedures are very successful indeed in identifying the best clustering alternatives when we cluster samples into homogeneous sets both in the univariate and the multivariate cases.

Moreover, the definition of MAICE gives a clear formulation of the principle of parsimony in statistical model building or comparison as the above examples demonstrate. And MAICE provides a versatile procedure for statistical model identification which is free from the ambiguities inherent in the application of conventional statistical procedures.

## Acknowledgement

This paper is a summary of some of the results in the Ph.D. thesis of the first author in the Department of Mathematics at the University of Illinois at Chicago, under the supervision of the second author. The first author is grateful to Professor Stanley L. Sclove for his valuable and useful comments.

# REFERENCES

[1] Ambramowitz, M., and Stegun, I.A. (1968), Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables (Nat. Bur. of Stand. Appl. Math. Ser., No. 55), 7th printing. U.S. Govt. Printing Office, Washington, D.C.

[2] Akaike, H. (1973), "Information Theory and an Extension of the Maximum Likelihood Principle," 2nd International Symposium on Information Theory, eds. B.N. Petrov and F. Csaki, Budapest: Akademiai Kiado, 267-281.

[3] ________ (1974), "A New Look at the Statistical Model Identification," IEEE Transactions on Automatic Control, AC-19, 716-723.

[4] Anderson, T.W., and Sclove, S.L. (1978), An Introduction to the Statistical Analysis of Data, Boston: Houghton Mifflin Company.

[5] Brown, B.W., Jr., and Hollander, M. (1977), Statistics: A Biomedical Introduction, New York: John Wiley.

[6] Chen, Hwei-Ju, Gnanadesikan, R., and Kettenring, J.R. (1974), "Statistical Methods for Grouping Corporations," Sankhya B, 36, 1-28.

[7] Delany, M.J., and Healy, M.J.R. (1966), "Variation in the White-toothed Shrews (Crocidura spp.) in the British Isles," Proceedings of the Royal Society, B, 164, 63-74.

[8] Dolkart, R.E., Halpern, B., and Perlman, J. (1971), "Comparison of Antibody Responses in Normal and Alloxan Diabetic Mice," Diabetes, 20, 162-167.

[9] Duran, B.S., and Odell, P.L. (1974), Cluster Analysis: A Survey, New York: Springer-Verlag.

[10] Fisher, R.A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, 7, 179-188.

[11] Friedman, H.P., and Rubin, J. (1967), "On Some Invariant Criteria for Grouping Data," Journal of the American Statistical Association, 62, 1159-1178.

[12] Heidorn, K.C. (1979), "Synoptic Weather Patterns Associated with Nitrates in Suspended Particulate in Southern Ontario," Water, Air, and Soil Pollution, 11, 225-235.

[13] Kendall, M.G. (1966), "Discrimination and Classification," in P.R. Krishnaiah (Ed.), Multivariate Analysis, New York: Academic Press.

[14] Mezzich, J.E., and Solomon, H. (1980), Taxonomy and Behavioral Science, New York: Academic Press.

[15] Solomon, H. (1971), Numerical Taxonomy, Mathematics in the Archaeological and Historical Sciences, Edinburgh: Edinburgh University Press, 62-81.

[16] Williams, W.H., and Goodman, M.L. (1971), "A Statistical Grouping of Corporations by Their Financial Characteristics," Journal of Financial and Quantitative Analysis, 1095-1104.

[17] Wilks, S.S. (1932), "Certain Generalizations in the Analysis of Variance," Biometrika, 24, 471-494.

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

**REPORT DOCUMENTATION PAGE**

READ INSTRUCTIONS
BEFORE COMPLETING FORM

1. REPORT NUMBER
Technical Report 82-1

2. GOVT ACCESSION NO.
AD-A112 769

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Multi-Sample Cluster Analysis Using Akaike's Information Criterion

5. TYPE OF REPORT & PERIOD COVERED
Technical Report

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Hamparsum Bozdogan and Stanley L. Sclove

8. CONTRACT OR GRANT NUMBER(s)
N00014-80-C-0408

9. PERFORMING ORGANIZATION NAME AND ADDRESS
University of Illinois at Chicago Circle
Box 4348, Chicago, IL 60680

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS

12. REPORT DATE
January 30, 1982

13. NUMBER OF PAGES
32+ii

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)
Office of Naval Research
Statistics and Probability Branch
Arlington, VA 22217

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Unlimited distribution

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Multi-sample cluster analysis; Akaike's Information Criterion (AIC); ANOVA Model; MANOVA Model; maximum likelihood

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Multi-sample cluster analysis, the problem of grouping samples, is studied from an information-theoretic viewpoint via Akaike's Information Criterion (AIC). This criterion combines the maximum value of the likelihood with the number of parameters used in achieving that value. The multi-sample cluster problem is defined, and AIC is developed for this problem. The form of AIC is derived in both univariate and multivariate analysis of variance models. Numerical examples are presented and results are shown

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

to demonstrate the utility of AIC in identifying the best clustering alternatives.

ILMED

8